Sequence analysis

Explore Positional Preference of Mutation Through Correlation Between Sequence Variations and Structural Changes

Mengfei Cao
Department of Computer Science

ABSTRACT

Motivation: Exploring evolution on proteins has always been an attractive opportunity but also a harsh challenge for scientists. Capturing the evolutionary information will help various computational biology tasks, such as structural prediction and functional prediction. By making use of both structural information and sequence information on proteins, we try to extract and quantitate useful evolutionary information, further improving the remote homology detection as our final purpose.

Results: Preliminary result is presented on superfamily Cytochrome_c following the SCOP hierarchy. We demonstrate the critical demand for positional preference in the homology detection routine using simulated evolution. In addition, even though we have not seen quite positive tendency that relationship between sequence changes and structural variations imposes direct effect performance combining simulated evolution, it is expected that with complete experimental results accomplished and larger scale of data experimented on, we would observe boosted consequence from the correlational analysis. The most important of all, this possible conclusion will give a revolutionary understanding in the different effect of amino acid mutations among protein primary structures and protein tertiary structures.

1 INTRODUCTION

A lot of work has been done around how to dig up the principles of mutation over protein amino acids (Socolich et al, 2005; Sasidharan and Chothia, 2007; Sadowski and Jones, 2009; Kumar and Cowen, 2009, 2010). In general, the information embedded among sequences and structures gives the constraints of mutability in terms of the location concerned with amino acids close to this location. A mutation will not corrupt the tertiary structures, nor will disrupt protein's functions. Matrices such as BLOSUM (Eddy, 2004) or PAM (Dayhoff, 1978) capturing the statistical pairwise patterns are commonly used as the pairwise substitution score measurement. More specifically, on beta-structural motifs residues that are hydrogen bonded in beta-sheets conserve high pairwise consistency(Kumar and Cowen 2010); on alpha-helical proteins, amino acids among paired coils are greatly correlated with their close neighbors(Berger, et al, 1995; McDonnell, et al, 2006). These principles can be used to capture the past mutational behavior and researchers attempted to dig them up through various methods such as Statistical Coupling Analysis(SCA), and probability approximation through co-occurrence frequencies.

A simple but excellent method (Sadowski and Jones, 2009) is brought forward so as to calculate the structural importance of amino acid positions by measuring the correlation between global structural change and the degree of mutational difference at a certain position within a multiple alignment. As is known widely, structure is more conserved than sequence. Therefore, the correlation between these two types of information should imply the gap between sequence mutations and structural conservations. In particularly, the less correlation a certain position reveals, the more likely this position has gone through mutations.

Correlational analysis can be easily embedded into this framework after constructing the feature vectors representing structural changes and vectors for local sequence changes. Global structural feature vectors can be simply obtained through RMSD and local sequence feature vectors can be calculated via pairwise score from substitution matrices, such as BLOSUM and PAM. Hereafter, correlational co-efficient is calculated between these two types of vectors at each position. Further, the achieved correlational values will indicate the preference of mutation and thus facilitate the simulated evolution models used in remote homology detection and various computational biology tasks.

Consider the following situation on a protein where the primary structure has gone through several mutations on some positions, however, its tertiary structure did not vary but conserved mostly the original 3-dimensional structure. Therefore, those mutated positions should have lower consistency with the global structure than those without mutations. One measurement that can be calculated to capture this consistency is to compare the sequence changes to the global structure variations among proteins over evolutions. It is expected that low correlation should reveal low consistency, and thus high likelihood of mutations. Motivated by this intuition, our work attempts to empirically verify the usage of correlational analysis in approximating the positional preference for mutation. Preliminary result is presented and brings both demand and hope to light.

2 METHODS

In order to quantitate the evolutionary information, we employed a method based on the analysis of correlation between evolutionary conservation at a sequence position and change to global tertiary structure (Sadowski and Jones, 2009). We made the conjecture that

*COMP-167 Final Proposal, 2011 Fall.

the lower the evolutionary conservation there is at a certain position, the higher the likelihood that the position has gone through mutations within remote homology space. Then motivated by simulated evolution model (Kumar and Cowen, 2009), we can apply this mutational preference into the prior knowledge as input to the simulated evolution model, resulting in more accurate and appropriate artificial sequences in the training stage for profile hidden Markov model. In particular, we are exploring the probability that a certain position should be mutated within simulated evolution model; this probability should be approximated as, or at least positively correlated with, the correlation between sequence variation and structural changes within the homology space. Therefore, by quantitatively calculating this correlation, we get the approximate mutation probability so as to facilitate the simulated evolution model.

Particularly, we quantitate the correlation between sequence information and structural information by using correlational analysis on each column among multiple structural/sequence alignment. Then we use this correlation as the positional preference for mutation.

Given *N* proteins within the same homology space, we conduct multiple structure alignment and obtain both structural alignment and sequence alignment with *Q* columns. For each column, we calculate the representation though difference between *M* pairs of proteins, where *M* equals *(N-1)*N/2*. Lastly we reach the positional preference for mutation.

More specifically, to generate the feature vectors representing global structural information for the alignment and each position's local sequence information for the N columns, the pipeline (Figure 1) is followed:

(1) The input to positional preference framework is a set of protein structural data within the same homology space.

(2) The proteins are structurally aligned using multiple structural alignment program, resulting in both multiple sequence alignment and multiple structure alignment.

(3) For each pair of proteins, pairwise RMSD (root-mean-square deviation) is computed on the superposition within the whole structural alignment (Figure 2).

(4) For each column of the sequence alignment, the pairwise sequence scores are calculated using the substitution matrix BLOSUM62.

(5) On each position represented by column, Pearson correlation co-efficient(PCC) can be calculated with respect to the pairwise RMSD, the measurement of global structural changes, and pairwise sequence score, the measurement of local sequence changes. Output PCCs on each column as the positional preference for residue mutation.

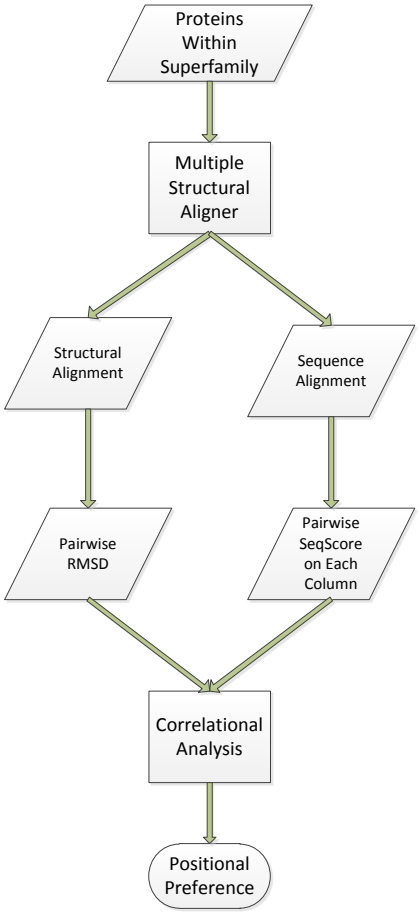

Fig. 1. Framework for Calculating Positional Preference.

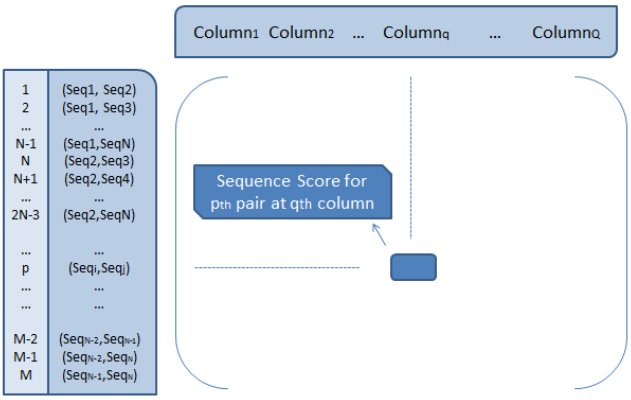

Fig. 2. Sequence Score Vector for Each Column Representing Sequence Changes

2.1 Datasets

Due to the final usage in remote homology detection on alpha-helical proteins, we will conduct the positional preference analysis on all training groups of superfamilies. In particular, after removal of redundancy among proteins in SCOP(Lo Conto et al., 2002), superfamilies that have at least three families consisting of at least 5 sequences are chosen as the experimental data. Among each superfamily, leave-one-family-out cross validation is conduct, in which the positional preference analysis for each training group of proteins is carried out. In details, on each round one family of proteins is extracted and reserved as the test data while the other families of proteins are used to train through multiple structure alignment, mutational analysis, simulated evolution, and profile hidden Markov modeling. Finally the reserved family in advance, together with equally sampled outliers are used to run the homology detection test.

2.2 Multiple Structural Alignment

In the positional preference framework, the module of multiple structure alignment is required and the required condition is that: the aligner should be able to give both structural alignment and sequence alignment. Above all, high evolutionary conservation and structural consistency are demanded. Therefore we employed one of the common used programs, MATT (Menke et al., 2008).

2.3 Global Structural Variations and Local Sequence Changes

Pairwise structural similarity scores are calculated using the RMSD on superpositions based on equaivalences taken from the multiple alignments; a fully pairwise RMSD vector for all pairs of proteins is thus used to represent the global structural variations.

$$RMSD(protein_1, protein_2) = \frac{\sum_i Eucli(\, atom_{1,i}, atom_{2,i})}{U},$$

where Eucli(*,*) is the Euclidean distance operator, $atom_{1,i}$ and $atom_{2,i}$ are the two C*a* atoms at i_{th} equivalent position in the alignment, U is the total number of columns without gaps for the two proteins in the alignment. This score is used as one entry of the global structural vector with all pairs of RMSD.

As for the sequence changes on each position, every residue pair in the alignment is scored using the substitution matrix BLOSUM62. Any pairs with gaps are scored as zero.

2.4 Correlational Analysis

The Pearson correlational coefficient (PCC) is calculated between global structural pairwise similarity score and each pairwise sequence score. The PCC at p_{th} column is computed as follows:

$$PCC_p = \frac{\sum_i (Stru_i - mStru)(SeqScore_{p,i} - mSeqScore_p)}{\sqrt{\sum_i (Stru_i - mStru)^2 \sum_i (SeqScore_{p,i} - mSeqScore_p)^2}}$$

where $Stru_i$ is the RMSD on i_{th} pair of proteins, *mStru* is the mean over all pairs, $SeqScore_{p,i}$ is the sequence score at p_{th} column on i_{th} pair, and $mSeqScore_p$ is the mean at p_{th} column.

Therefore, for each column we can get a *PCC*, used as the positional preference for mutation. Specifically, if we obtain |*PCC*| with value close enough to *1*, we may be pretty sure that since this position is totally consistent with the tertiary structure and thus very unlikely there has been some special mutations at this position, without considering the case where highly consistent mutation pair exists. At the other hand, if we achieve a *PCC* much close to *0*, we should be confident that since the residues at this position are extremely inconsistent with the structural changes, either structure has gone through significant changes or this position has gone through various mutations, unless the alignment is falsely formed with proteins in different homology spaces.

3 RESULTS

In preliminary experiments, we have test on superfamily Cytochrome_c, and its three leave-one-family-out training groups of proteins for each family: Monodomain Cytochrome_c, N-terminal domain, and Cytochrome bc1 domain. The reason I choose this superfamily firstly is that on the remote homology detection experiments, this superfamily yields the most degradation from detection with simulated evolution to that without simulated evolution. Specifically, Table 1 shows the number of columns in alignments, the number of proteins in each training data, and performance of detection using simulated evolution (the 4th column) and the performance without simulated evolution. As indicated, on family Monodomain C., the detection AUC (area under the ROC curve) degrades from 88.55% to 66.49%.

Family Name	#columns	#Proteins	AUC(SimEvo)	AUC
Monodomain C.	3250	24	66.49%	88.55%
N-terminal	714	88	100%	100%
Cytochrome bc1	714	88	88.89%	86.11%

Table 1. Related Descriptions of Superfamily Cytochrome_c

3.1 Column-wise Correlation

I calculated the PCCs for columns without gaps, among which there are 3 columns in the data on family N-terminal domain, where all sequences have the same residue and thus there is no variation. The limit of *PCC* in these cases is 1 and intuitively those columns should be considered as the most conserved positions and the likelihood of mutation should be the smallest. These cases are rare; however, what might be of interest is that these cases only exist in two families data: N-terminal and Cytochrome bc1, where simulated evolution didn't degrade. It might reveal that on these two families proteins are more convergent and simulated evolution helps expand the homology space appropriately, while on family Monodomain, proteins are diverse and it is likely that augmented data from simulated evolution pulls the training data out of the homology space, thus degrading the detection performance.

Monodomain C.		N-terminal	
column	|PCC|	column	|PCC|
2068	0.3590	340	0.3701
1224	0.1644	339	0.3162
1243	0.1543	400	0.3099
2067	0.1364	337	0.2833
1247	0.1314	255	0.2783

Table 2. Top 5 Absolute PCCs

Table 2 gives 5 columns with highest absolute PCCs on two families training data. In addition, there are in total 13 columns on family Monodomain C., where there is no gap; while there are 53 such columns on family N-terminal. This result is consistent on the former observation that proteins on family N-terminal's training data are more convergent and thus augmented data from simulated evolution would help with higher likelihood.

However, from the values of PCCs, family N-terminal tends to have more columns with higher absolute PCCs considering its total columns as 714 comparing to 3259 on family Monodomain C.. It is likely that more positions among family N-terminal's alignment have higher PCCs, and thus these positions indicate high correlation between sequence changes and structural variations. It goes against our conjecture that high correlation reveals low likelihood of mutation if the overall tendency also follow this pattern, because on family N-terminal we obtained relatively better detection performance with simulated evolution. Yet it is also possible that these positions with high PCCs have so singular mutation probability distribution that in the augmented data these positions tend to keep the original residue instead of changing due to high conservation. Also, we might also need more sophisticated observations on the current results. Therefore next we look further into the distributions of columns in terms of varieties.

In order to look more specifically on the columns with high correlation values and the columns with low correlation values, Figure 3 & 4 show the column 2068 with its 3 neighbors(col_2066, col_2067, col_2068, col_2069 from left to right) as well as their absolute PCCs and the column 1225 which has lowest absolute PCC as well as its one neighbor(col_1224, col_1225, from left to right) on family Monodomain C.'s training data,. Moreover, the entropies are calculated on the two columns distributions, and the entropy on column 2068 is bigger than that on column 1225. Therefore, higher PCC doesn't indicate lower varieties or lower varieties can't guarantee high PCC, since here column 2068 has both higher PCC and higher entropy.

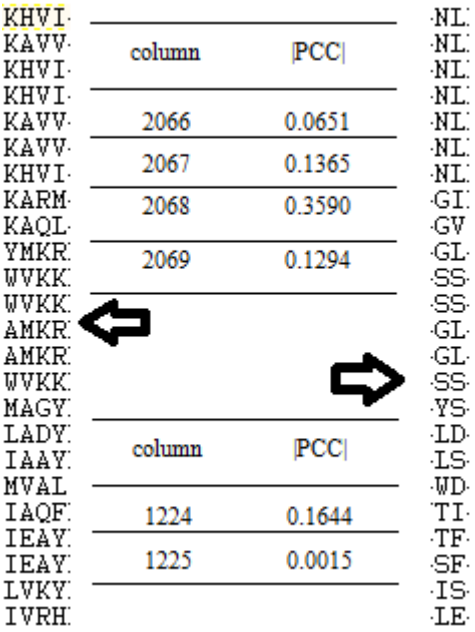

Fig. 3. Residues on Column with Highest Absolute PCC and that with Lowest Absolute PCC

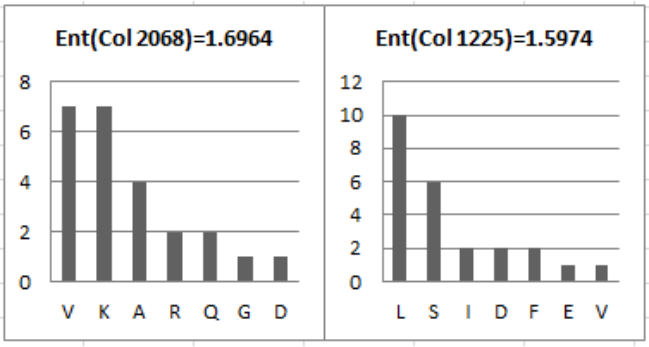

Fig. 4. Distribution of Residues on Column with Highest Absolute PCC and Column with Lowest Absolute PCC (Ent stands for the entropy of the distribution)

3.2 *Comparing the Overall Correlation

Here I also calculate the overall pairwise sequence score on the whole alignment. Then the Pearson correlation coefficients are calculated; results -0.1323 and -0.3556 are obtained on family Monod-omain C. and family N-terminal respectively. The family that doesn't degrade from simulated evolution has higher absolute correlation coefficient. This gives similar prompt as before that goes against our conjecture; however, this fact probably comes from the difference of diversities. On the other hand, it also implies that positional preference analysis on individual position is required instead of the overall decision on all positions

Intuitively, low variety might increase correlations between sequence changes and structural variations because proteins are conserved on both sequence and structure. Thus it makes sense that 1)

experimentally family N-terminal should have low correlation values but due to its low variety 2) it ends up with higher correlations on the conserved columns. For the rest columns, the less conserved ones, it is conceivable that they should have lower correlation values and needs mutation, which explains the non-degraded performance of homology detection with simulated evolution.

3.3 To Do List

Probably the best way to untangle previous confusion is to complete the experiment by conducting simulated evolution and taking the *(1-PCC)* as the input prior mutation probability. Hereafter apply profile training program and obtain the profile hidden Markov model. Consequently employ the reserved test data on HMM, resulting in the homology detection performance, which might give the answer whether or not the positional preference characterize the mutation behavior accurately. Particularly, if the homology detection performance with positional preference yields better result, it can be accepted that the correlational analysis do give appropriate descriptions regarding the sequence mutation and preserving the structural conservation.

Above is the complement to the preliminary experiment on superfamily Cytochrome_c. More experiments should be done on more superfamilies in the homology detection test including the other 10 superfamilies of 66 families' leave-one-family-out cross validation routine. For each of the experiment, the following routine is carried out:

- Firstly, multiple structure alignment is conduct using some multiple structural aligner;
- Secondly, correlational analysis is executed on the alignment so as to obtain the mutational preference on each position;
- Thirdly, simulated evolution model is stimulated together with the positional preference for mutations; in particularly in this stage, protein sequences are augmented with artificial sequences. These new generated sequences come from simulating mutation by approximating the residue mutation probability distribution;
- Further, a profile HMM is trained from both the original real sequence and artificially generated sequences, and then homology detection stage is on with the reserved testing homology proteins and sampled outliers. The area under the curve (AUC) of receiver operating characteristic(ROC) is used as the target to measure the accuracy and describability of the whole model on real world evolution. By comparing the detection performance to that without positional preference for mutation, it can be intuitionally determined whether or not this mutational analysis is appropriate.

4 DISCUSSION

The correlational analysis for mutation makes it possible to quantitate correlation between sequence information and structure information. As has been accepted, structure is more conserved than sequence and thus from this gap we hope to handle the probability that an amino acid position has gone through mutations within homology space.

The preliminary results show that more convergent group of proteins yields better homology detection if augmented data from simulated evolution is utilized. More importantly, the results show that: a) on the training data of family Monodomain, since we know this dataset is quite diverse and thus the neighborhood of each position gives pretty noisy information on approximating the mutation distribution during simulated evolution, resulting in false decisions in mutating; b) There are positions with high correlation values but also high varieties. Thus we should make use of the positional analysis so as to reduce the mutation on positions with high correlation values because they preserve the characteristics of homology space.

More experiments remain to be conduct on the other 66 families' training data. Ideally, the performance of homology detection can be boosted by incorporating the positional preference when executing mutations.

REFERENCES

Berger,B. et al. (1995) Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci.* USA, 92, 8259–8263.

Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. (1978). "A model of Evolutionary Change in Proteins". *Atlas of protein sequence and structure (volume 5, supplement 3 ed.). Nat. Biomed. Res. Found*.. pp. 345-58.

Eddy,S. (2004) Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnol*., 22, 1035.

Henikoff, S.; Henikoff, J.G. (1992). Amino Acid Substitution Matrices from Protein Blocks. *PNAS 89* (22).

Kumar,A. and Cowen,L. (2009) Augmented training of hidden Markov models to recognize remote homologs via simulated evolution. *Bioinformatics*, 25, 1602-1608.

Kumar,A. and Cowen,L. (2010) Recognition of beta structural motifs using hidden Markov models trained with simulated evolution. *Bioinformatics*, 26:i287{i293}.

Lo Conte,L. *et al.* (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acid Res*., **30**, 264–267.

McDonnell AV, Jiang T, Keating AE, Berger B (2006): Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*,22:356-358.

Menke,M. et al. (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol*., 4, 88-99

Rost, B., and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct. Funct. Genet*. 19, 55-72

Sadowski, MI., Jones DT (2009) An automatic method for assessing structural importance of amino acid positions. *BMC Struct Biol*, 9:10..

Sasidharan R, Chothia C (2007): The selection of acceptable protein mutations. *Proc Natl Acad Sci USA*, 104:10080-10085.

Socolich, M. et al. (2005) Evolutionary information for specifying a protein fold. *Nature doi:10.1038/nature03991*

Valencia,A. and Pazos,F. (2003) Prediction of protein-protein interactions from evolutionary information. *In Bourne,P.E.and Weissig,H. (eds), Structural Bioinformatics. Wiley Inc.*, pp. 411-426.